Approximating k-Median with Non-Uniform Capacities

Julia Chuzhoy* Yuval Rabani†

July 3, 2003

Abstract

In this paper we give a constant factor approximation algorithm for the capacitated k-median problem. Our algorithm produces a solution where capacities are exceeded by at most a constant factor, while the number of open facilities is at most k. This problem resisted attempts to apply the plethora of methods designed for the uncapacitated case. Our algorithm is based on adding some new ingredients to the approach using the primal-dual schema and lagrangian relaxations.

Previous results on the capacitated k-median problem gave approximations where the number of facilities is exceeded by some constant factor. Relaxing the constraint on the number of facilities seems to render k-median problems much simpler. In some applications it is important not to violate the constraint on the number of facilities, whereas relaxing the capacity constraints is a natural thing to do, as the capacities express rough estimates on cluster sizes.

*Computer Science Department, Technion — IIT, Haifa 32000, Israel. Email: cjulia@cs.technion.ac.il

†Computer Science Department, Technion — IIT, Haifa 32000, Israel. Work supported by BSF grant number 99-00217, by ISF grant number 386/99, by IST contract number 32007 (APPOL), and by the Fund for the Promotion of Research at the Technion. Email: rabani@cs.technion.ac.il

1 Introduction

In this paper we consider the following *capacitated k-median* problem. The input is a set of clients C, a set of potential facilities F, a non-negative integer capacity function $u : F \to \mathbb{N}$, a metric distance d on $C \cup F$, and a positive integer k. The desired output is a subset of open facilities $F' \subset F$ of cardinality at most k and an assignment of clients to open facilities $\varphi : C \to F'$ such that for all $i \in F'$, $|\varphi^{-1}(i)| \leq u(i)$. The objective is to minimize the total assignment cost $\sum_{j \in C} d(j, \varphi(j))$. In the related *capacitated facility location* problem the input includes non-negative facility costs f_i for all $i \in F$. The output set of facilities F' may be of any positive cardinality, and the objective is to minimize the total assignment cost plus the total opening cost $\sum_{j \in C} d(j, \varphi(j)) + \sum_{i \in F'} f_i$.

Facility location and k-median problems are motivated by applications in logistics, distributed systems, clustering, and other areas. In recent years variants of these problems (mostly the uncapacitated versions) inspired an explosion of research promoting diverse methods. Constant factor approximations were discovered for uncapacitated facility location [20] and for uncapacitated k-median [4] (both used the LP relaxation of [7] and the filtering technique of [14]). The approximation factors for both problems were improved significantly using the original filtering/LP-rounding ideas [5, 6, 9], the primal-dual schema and lagrangian relaxations [12, 3], local search heuristics [13, 1], and dual fitting [15, 11, 10, 21, 16]. (See also the surveys in [8, 19].) The current approximation guarantee champions for these problems [1, 16] are quite close to the known hardness of approximation results. Constant factor approximations were discovered for the capacitated facility location problem [12, 18].

Unfortunately, none of this impressive collection of results seemed to apply directly to the capacitated k-median problem. Standard LP relaxations are known to have unbounded integrality gap for this problem [4]. If the capacity constraints are relaxed, and the approximation algorithm is allowed to exceed the capacities by a constant factor, this is no longer necessarily true. However, the methods used to solve uncapacitated k-median all seem to suffer from serious drawbacks when trying to apply them to the capacitated problem. For example, capacitated facility location can be solved using the primal-dual schema [12]. However, this is done by reducing the problem to the uncapacitated case, changing the metric to include some of the facility opening costs in the process. The resulting bounds are not of the form that can be used in the lagrangian relaxation approach that solved the uncapacitated case. The only previous attempts to address the capacitated k-median problem (to the best of our knowledge) got a constant factor approximation with a constant factor violation of the capacities for the case of uniform capacities [4], or while exceeding the number of facilities k by a constant factor [2]. Solving the uniform capacities case requires a simple modification of the uncapacitated algorithm. Relaxing the restriction on the number of facilities usually renders k-median problems much easier.

In this paper we give a constant factor approximation algorithm for the capacitated k-median problem. Our algorithm produces a solution where capacities are exceeded by at most a constant factor. We note that in clustering applications where the capacities are estimates on cluster sizes, relaxing the constraints imposed by those estimates might be a reasonable thing to do. Our algorithm is based on a simple primal-dual algorithm for capacitated facility location that also exceeds the capacities by constant factors, but on the

other hand produces solutions that can be used in the lagrangian relaxation approach. Our solution differs from previous results (e.g. [12]) in two main respects. Firstly, we base our primal-dual algorithm not on the usual LP relaxation of [7], but on a different relaxation similar to the one used in dual fitting. Secondly, our procedure that generates a k-median solution is much more complicated than similar procedures in previous work, and its analysis has to amortize the cost of some connections against others in a non-trivial fashion. The reason for this is that some clients have to be rerouted to far-away facilities to accommodate the (relaxed) capacity constraints. We note that in this extended abstract, we made no attempt to optimize the constant guarantees.

2 Preliminaries

Suppose we are given some solution to a capacitated facility location problem. Let Φ denote the client connection cost and Ψ denote the facility cost of the solution. Our algorithm for the capacitated k-median problem can be based on any approximation algorithm for the capacitated facility location problem, which always produces a solution where the capacities are violated by at most a constant factor, and the inequality $\Phi + r\Psi \leq rOPT$ holds for some constant r, where OPT is the value of the optimal solution. We remark that the algorithm of Jain and Vazirani [12] produces a solution of the desired form for the uncapacitated facility location problem. However, this does not seem to extend to arbitrary capacities. Therefore, we show a new approximation algorithm for the capacitated facility location problem with the above properties, which is based on LP relaxation similar to the one used in [10], rather than the usual LP relaxation of [7].

Our algorithms are motivated by and analyzed using the following integer linear program formulation of the capacitated facility location problem. To state the formulation, we use the following notation: For every facility i and every subset of clients D of size $|D| \leq u(i)$, we define a *star* $S = S(i, D)$. The cost of this star is $w(S) = f_i + \sum_{j \in D} d(i, j)$. We denote $f(S) = i$, $c(S) = D$, and $n(S) = |D|$. Let $\mathcal{S}$ be the set of all such stars. The integer linear program formulation is:

$$\begin{array}{lll} \text{minimize} & \sum_{S \in \mathcal{S}} w(S) z_S & \\ \text{s.t.} & & \\ & \sum_{S | j \in c(S)} z_S \geq 1 & \forall j \in C \\ & z_S \in \{0, 1\} & \forall S \in \mathcal{S} \end{array}$$

By relaxing the integrality constraints to $z_S \geq 0 \ \forall S \in \mathcal{S}$, we get a linear programming relaxation for the capacitated facility location problem. We denote this relaxation by (FLR). Notice that (FLR) has an exponential number of variables. However, we will not need to solve it. The dual LP, which we denote by (DFL), is:

$$\begin{array}{lll} \text{maximize} & \sum_{j \in C} \alpha_j & \\ \text{s.t.} & & \\ & \sum_{j \in c(S)} \alpha_j \leq w(S) & \forall S \in \mathcal{S} \\ & \alpha_j \geq 0 & \forall j \in C \end{array}$$

Let z^*_{fl} denote the value of an optimal solution for the capacitated facility location problem. Then by weak duality $z^*_{fl} \geq \sum_j \alpha_j$ for any dual feasible solution α.

Let z be an integral solution to (FLR). Note that it is possible that there are several stars with the same facility i in z. It is sometimes convenient to assume that we have n available copies of each facility, where each copy is viewed as a different facility. In this representation, each facility is opened at most once in z, and we can express z as a pair of integral vectors (x, y) in the following way. For every $i \in F$, $y_i = 1$ iff there is a star S with $f(S) = i$ and $z_S = 1$, and for every $i \in F$ and $j \in C$, $x_{ij} = 1$ iff there is a star S with $f(S) = i$, $j \in c(S)$, and $z_S = 1$. We denote by $\Phi(z) = \Phi(x, y) = \sum_{i \in F,\ j \in C} d(i, j)x_{ij}$ (the total connection cost of z), and we denote by $\Psi(z) = \Psi(x, y) = \sum_{i \in F} f_i y_i$ (the total opening cost of z).

We will use the following mixed integer program formulation of the capacitated k-median problem

$$
\begin{array}{lll}
\text{minimize} & \sum_{j \in C} \sum_{i \in F} d(i, j)x_{ij} & \\
\text{s.t.} & & \\
 & \sum_{i \in F} x_{ij} \geq 1 & \forall j \in C \\
 & \sum_{i \in F} y_i \leq k & \\
 & x_{ij} \leq y_i & \forall i \in F,\ \forall j \in C \\
 & y_i \in \{0, 1\} & \forall i \in F
\end{array}
$$

We denote by (KMR) the linear programming relaxation derived by relaxing the integrality constraints to $y_i \geq 0\ \forall i \in F$. We denote by z^*_{km} the value of an optimal solution for the capacitated k-median problem.

In both the capacitated k-median problem and the capacitated facility location problem we are interested in infeasible solutions that may connect too many clients to a facility. However, we restrict the excess to at most γ times the capacity, for some constant $\gamma > 1$. More formally, if φ is the assignment function of clients to facilities in our solution, we require that $|\varphi^{-1}(i)| \leq \gamma u(i)$ for all facilities i. We call such a solution a γ-*feasible* solution. (Similarly, a fractional solution is γ-feasible iff it violates the capacity constraints by no more than a factor of γ.) We call an algorithm that produces a γ-feasible solution whose cost is at most β times the optimum a (β, γ)-*approximation*.

3 Facility Location

In this section we present and analyze our bi-criteria approximation algorithm for the facility location problem. Our algorithm will be used in the approximation of the capacitated k-median problem. The algorithm is made up of two phases. We define and discuss each phase separately.

Phase 1.

The goal of this phase is to find a feasible solution to the dual program (DFL). This solution will serve as a lower bound on the cost of the optimal solution. We also obtain an infeasible primal solution. In the second phase, this solution will be converted into a solution satisfying the relaxed capacity constraints.

We start with all dual variables equal to 0. All the facilities are closed, and no stars are allocated. For each facility i, in this phase, we are going to allocate at most one star whose center is i. As soon as such a star is allocated, the facility i is declared "open". Allocating stars and opening facilities is done as follows. While there are unconnected clients, increase uniformly the values α_j for all unconnected clients j, until one of the following conditions holds:

1. There is a currently closed facility i and a set D of clients (connected or unconnected) of size at most $u(i)$, such that $\sum_{j \in D}(\alpha_j - d(i,j)) = f_i$. If the condition holds for some facility i and set D of clients, we allocate the star $S = S(i,D)$. Facility i is declared open. All the clients in $c(S)$ that are currently unconnected become connected to i. If, for some unconnected client $j \notin c(S)$, $\alpha_j \geq d(i,j)$ holds, then we connect j to i.

2. For some unconnected client j and open facility i, α_j becomes equal to $d(i,j)$. Then we connect j to i.

Notice that in the end all the clients are connected. However, the number of clients connected to i may thus exceed $u(i)$. This will be fixed in phase 2.

Claim 1. If client j is connected to facility i in the end of phase 1, then $\alpha_j \geq d(i,j)$. Also, if $j \in c(S)$ for some star S allocated by the algorithm, then $\alpha_j \geq d(j, f(S))$ (note that in this case j is not necessarily connected to $f(S)$).

Proof: First, assume $j \in c(S)$ for some star S allocated by the algorithm, with $f(S) = i$. Then,

$$\sum_{j' \in c(S) \setminus \{j\}} \alpha_{j'} \leq f_i + \sum_{j' \in c(S) \setminus \{j\}} d(i,j')$$

(otherwise the first condition was true for the star $S' = S(i, c(S) \setminus \{j\})$ before S was allocated). Since the first condition holds for S, $\alpha_j \geq d(i,j)$.

Now suppose j is connected to some facility i, a star S with $f(S) = i$ is allocated by the algorithm, and $j \notin c(S)$. Then if j connected to i when star S was allocated, $\alpha_j \geq d(i,j)$ clearly holds. If j connected to i later, then $\alpha_j = d(i,j)$. □

Claim 2. Phase 1 can be implemented to run in polynomial time.

Proof: The number of iterations is bounded by $|C|$. We show that each iteration can be implemented to run in polynomial time. Consider an iteration. Let A be the set of currently open facilities, and let $B = F \setminus A$. Let C_1 be the set of currently connected clients, $C_2 = C \setminus C_1$. Our goal is to compute the increase Δ in values of α_j for all $j \in C_2$ in this iteration, and in case the first rule must be applied, we must find the star that should be allocated.

Consider a facility $i \in B$. For each $0 \leq r_1 \leq |C_1|$, $0 \leq r_2 \leq |C_2|$, such that $r_1 + r_2 \leq u_i$, let $D(i, r_1, r_2)$ be the set of r_1 connected clients and r_2 unconnected clients that minimizes

$$\delta(i, r_1, r_2) = f_i + \sum_{j \in D(i,r_1,r_2)} d(i,j) - \sum_{j \in D(i,r_1,r_2)} \alpha_j.$$

Clearly, set $D(i, r_1, r_2)$ consists of r_1 connected clients and r_2 unconnected clients with the largest $(\alpha_j - d(i,j))$ values. Notice that the first condition holds at the end of current iteration for some star S with $f(S) = i$, and $c(S)$ a set of r_1 connected and r_2 unconnected clients if and only if $\Delta = \frac{\delta(i,r_1,r_2)}{r_2}$.

Let Δ_1 be the minimum value of $\frac{\delta(i,r_1,r_2)}{r_2}$ over all facilities $i \in B$, and all integers $0 \leq r_1 \leq |C_1|$, $0 \leq r_2 \leq |C_2|$, such that $r_1 + r_2 \leq u_i$. The first condition holds for some star in current iteration if and only if $\Delta = \Delta_1$. Put $\Delta_2 = \min_{j \in C_2, i \in A}\{d(i,j) - \alpha_j\}$. Clearly, the second condition holds in this iteration if and only if $\Delta = \Delta_2$. Thus, the increase in values α_j for $j \in C_2$ in current iteration is $\Delta = \min\{\Delta_1, \Delta_2\}$, and can be computed in polynomial time. In case $\Delta = \Delta_1$, the first condition holds. In this case we allocate the star $S(i, D(i, r_1, r_2))$, $i \in B$ for which $\frac{\delta(i,r_1,r_2)}{r_2} = \Delta$. □

Claim 3. The vector α computed in phase 1 is a dual feasible solution.

Proof: Consider a star S. Let $i = f(S)$. If i is closed, then $\sum_{j \in c(S)} \alpha_j \leq w(S)$, for otherwise S or some other star with i as its center would have been opened by the algorithm. Suppose i is open at the end of the algorithm, and S' with $f(S') = i$ is allocated. If $S' = S$, then trivially $\sum_{j \in S} \alpha_j = w(S)$, because as soon as this condition holds and S is allocated, all $j \in S$ that were not connected previously, get connected and their α_j-s stop increasing. If $S' \neq S$, partition $c(S)$ into two subsets. Let C_1 be the set of all clients $j \in S$ that had $\alpha_j \geq d(i,j)$ at the time when S' was allocated, and let C_2 be the set of all other clients in $c(S)$. When S' was allocated, the following condition was true: $\sum_{j \in C_1}(\alpha_j - d(i,j)) \leq f_i$. As for the clients $j \in C_2$, when phase 1 terminates $\alpha_j \leq d(i,j)$, because if at some point $\alpha_j = d(i,j)$, we connect j to i and stop increasing α_j. Therefore, at the end of phase 1, $\sum_{j \in S} \alpha_j \leq w(S)$. Thus, in all cases the dual constraints are satisfied. □

Phase 2.

In this phase we convert the solution produced in phase 1 into a solution satisfying the relaxed capacity constraints.

Put $\mathcal{S}' = \emptyset$. Given a star S with facility i, let $r(S) = \frac{f_i}{n(S)}$. Sort the stars allocated in phase 1 by non-decreasing $r(S)$ value. Scan the stars in this order. Let S be the current star. If there is no star $S' \in \mathcal{S}'$ such that $c(S') \cap c(S) \neq \emptyset$, add S to $\mathcal{S}'$. Let F' be the set of facilities i such that there is a star with center i in $\mathcal{S}'$. Note that for each client j, there is at most one star $S \in \mathcal{S}'$ that contains j. If there is a star $S \in \mathcal{S}'$ such that $j \in c(S)$, connect j to $f(S)$. We say that this is a *direct connection*. Let j be one of the remaining clients, and suppose that at the end of phase 1, j was connected to facility i. If $i \in F'$, then connect j to i. This is also called a direct connection. Otherwise, there is a client j' and a star $S' \in \mathcal{S}'$ such that $j' \in c(S) \cap c(S')$. Connect j to $f(S')$. We say that this is an *indirect connection* for which j' is *responsible*. Finally, for every $S \in \mathcal{S}'$ consider the set of clients D connected to $i = f(S)$. Open $\lceil |D|/5u(i) \rceil$ copies of i. Connect to each copy at most $5u(i)$ clients from D.

We first bound the connection cost.

Claim 4. At the end of phase 2, if client j is connected directly to facility i, then $d(i,j) \leq \alpha_j$, and if j is connected indirectly to i, then $d(i,j) \leq 3\alpha_j$.

Proof: The first part of the claim follows from Claim 1. As for the second part, suppose j is connected indirectly to i', and let j' be the client responsible for this connection. Let $S' \in \mathcal{S}'$ be the star with $f(S') = i'$. Let i be the facility to which j was connected at the end of phase 1, and let S be the star allocated in phase 1 with $f(S) = i$. From Claim 1, $d(i,j') \leq \alpha_{j'}$, $d(i',j') \leq \alpha_{j'}$ and $d(i,j) \leq \alpha_j$. We show that $\alpha_{j'} \leq \alpha_j$: When star S was allocated, either j' connected to i, or it was already connected to some other facility. Therefore, $\alpha_{j'}$ stopped growing by the time S was allocated. However, j connected to i either when it was allocated or later. Therefore, $\alpha_{j'} \leq \alpha_j$, and the connection cost $d(i',j) \leq d(i,j) + d(i,j') + d(i',j') \leq \alpha_j + 2\alpha_{j'} \leq 3\alpha_j$. □

Claim 5. Let $S \in \mathcal{S}'$, and let $i = f(S) \in F'$. We denote by $c'(S)$ the set of clients connected to i that do not belong to $c(S)$. Then the total cost of all the copies of i is at most

$$\sum_{j \in c(S)} (\alpha_j - d(i,j)) + \sum_{j \in c'(S)} \frac{\alpha_j}{4}.$$

Proof: As S was allocated in phase 1, we have $\sum_{j \in c(S)} \alpha_j = w(S)$. More than one copy of i is opened only if there are at least $4u(i)$ clients not in $c(S)$ connected to i. Pick $4u(i)$ of these clients and charge each of them with $\frac{f_i}{4u(i)}$. If t copies of i are opened, there are at least $5u(i)(t-2)$ uncharged clients not in $c(S)$. Charge each of these clients with $\frac{f_i}{5u(i)}$.

Consider a client $j \notin c(S)$ connected to i. Observe that S can only be allocated after time $\frac{f_i}{n(S)}$ in phase 1. If j is connected directly to i, then it was connected to i at the end of phase 1. Therefore, it connected to i at time $\frac{f_i}{n(S)}$ or later, and $\alpha_j \geq \frac{f_i}{n(S)} \geq \frac{f_i}{u(i)}$. If j is connected to i indirectly, then there is a star S' and $i' = f(S')$, such that j was connected to i' at the end of phase 1. By the order in which stars were added to $\mathcal{S}'$, $\frac{f_{i'}}{n(S')} \geq \frac{f_i}{n(S)}$. Therefore, $\alpha_j \geq \frac{f_{i'}}{n(S')} \geq \frac{f_i}{n(S)} \geq \frac{f_i}{u(i)}$. □

The main result of this section is the following bound on the performance guarantee of the algorithm.

Lemma 6. Let (x,y) be the solution computed by the above two-phase algorithm. Then,

$$\Phi(x,y) + 4\Psi(x,y) \leq 4 \sum_j \alpha_j \leq 4z^*_{fl}.$$

Proof: Let B be the set of directly connected clients belonging to stars in $\mathcal{S}'$, let D be the set of all the other directly connected clients, and let I be the set of indirectly connected clients. Then

$$\begin{aligned} 4\Psi(x,y) + \Phi(x,y) &\leq 4\Psi(x,y) + 4\sum_{j \in B} d(\varphi(j),j) + \sum_{j \in D} d(\varphi(j),j) + \sum_{j \in I} d(\varphi(j),j) \\ &\leq 4\sum_{j \in B} \alpha_j + 4\sum_{j \in D \cup I} \frac{\alpha}{4} + \sum_{j \in D} \alpha_j + 3\sum_{j \in I} \alpha_j \end{aligned}$$

$$\leq \; 4 \sum_{j \in C} \alpha_j \qquad \square$$

4 *k*-Median

Consider any capacitated facility location algorithm that produces a γ-feasible solution (x, y) such that $\Phi(x, y) + \beta\Psi(x, y) \leq \beta z^*_{fl}$. We have:

Lemma 7. The above-mentioned algorithm can be used to generate in polynomial time two γ-feasible solutions $S_1 = (x^1, y^1)$ with k_1 open facilities and $S_2 = (x^2, y^2)$ with k_2 open facilities, such that the following conditions hold:

1. $k_1 \leq k \leq k_2$.
2. $\frac{k_2 - k}{k_2 - k_1}\Phi(S_1) + \frac{k_1 - k}{k_2 - k_1}\Phi(S_2) \leq \beta \cdot z^*_{km}$. $\square$

The proof of this lemma is identical to the lagrangian relaxation argument of [12]. Alternatively, one can prove this lemma using a packing argument similar to the one in [17]. We do not include a proof in this extended abstract.

Thus, for the remainder of this section we assume that we are given two such solutions $S_1 = (x^1, y^1)$ and $S_2 = (x^2, y^2)$, generated using the algorithm from the previous section. Let A denote the set of open facilities in S_1, and let B denote the set of open facilities in S_2. For each client $j \in C$, let $\varphi_1(j)$ be the facility that serves j in S_1, and let $\varphi_2(j)$ be the facility that serves j in S_2.

Put $a = \frac{k_2 - k}{k_2 - k_1}$ and $b = \frac{k - k_1}{k_2 - k_1}$. The fractional solution $S = aS_1 + bS_2$ is a 5-feasible solution to (KMR). The solution cost is $a\Phi(S_1) + b\Phi(S_2) \leq 4z^*_{km}$. If $a \geq 0.1$, we can take the solution S_1. As $k_1 \leq k$, this is a 5-feasible solution for the k-median problem. Its cost $\Phi(S_1) \leq 10\Phi(S) \leq 40 z^*_{km}$.

Otherwise, we do the following. Let $U_1(i)$ denote the set of clients connected to facility $i \in A$ in solution S_1. Similarly, let $U_2(i)$ denote the set of clients connected to $i \in B$ in solution S_2. For each $i \in A$, $i' \in B$, we define $w_i(i') = \frac{|U_1(i) \cap U_2(i')|}{|U_1(i)|}$. Note that $w_i(i')$ is the fraction of clients in $U_1(i)$ that are connected to i' in solution S_2. Thus, for every $i \in A$, $\sum_{i' \in B} w_i(i') = 1$. We open (at random) a set $F = F_1 \cup F_2 \cup F_3$ of facilities. We will show that the number of open facilities is at most k with high probability. The set F is determined as follows.

Let $A_1 = \{i \in A \mid \exists i' \in B \text{ with } w_i(i') \geq 0.1\}$, and put $k' = |A_1|$. Clearly, $k' \leq k_1 < k$. Take $F_1 \subseteq B$ to be a set of size k' such that for every $i \in A_1$, there is at least one $i' \in F_1$ with $w_i(i') \geq 0.1$. There are $k_2 - k'$ facilities in $B \setminus F_1$ and $k_1 - k'$ facilities in $A \setminus A_1$. Notice that $a(k_1 - k') + b(k_2 - k') = k - k'$. Take F_2 to be a set of size $\lceil b(k_2 - k') \rceil$ chosen uniformly at random from $B \setminus F_1$. Next, consider $i \in A \setminus A_1$. Let $\{i_1, i_2, \ldots\}$ be the set of all facilities $i' \in B$ with non-zero weight $w_i(i')$, ordered by non-decreasing order of their distance from i. Put $t = t(i) = \max\{j \mid \sum_{s<j} w_i(i_s) < 0.5\}$, and put $d(i) = d(i, i_t)$. Put $B' = B'(i) = (F_1 \cup F_2) \cap \{i_1, \ldots, i_t\}$. If $\sum_{i' \in B'} w_i(i') < 0.1$, we add i to F_3. (Notice that $F_3 \subset A$, unlike F_1, F_2.) Finally, put $F = F_1 \cup F_2 \cup F_3$.

Claim 8. For every $i \in A \setminus A_1$, $Pr[i \in F_3] \leq \frac{a}{2}$.

Corollary 9. With probability at least $\frac{1}{2}$, $|F| = |F_1| + |F_2| + |F_3| \leq k$.

Proof: By Claim 8 and linearity of expectation, $E[|F_3|] \leq a(k_1 - k')/2$. Thus, by Markov's inequality, $\Pr[|F_3| > a(k_1 - k')] \leq \frac{1}{2}$. As the number of open facilities is integral, $|F_3| \leq \lfloor a(k_1 - k') \rfloor$ with probability at least $\frac{1}{2}$. Therefore, with probability at least $\frac{1}{2}$, $|F| = |F_1| + |F_2| + |F_3| \leq k' + \lceil b(k_2 - k') \rceil + \lfloor a(k_1 - k') \rfloor = k$. □

Proof of Claim 8: Consider the facilities $i_1, i_2, \ldots, i_t$. Let $p_{i'}$ denote the probability that $i' \notin F_1 \cup F_2$. Then, for $i' \in B$, $p_{i'} = 0$ if $i' \in F_1$ and $p_{i'} \leq 1 - b = a$ otherwise. For $j = 1, 2, \ldots, t$ define a random variable X_j. If $i_j \notin F_1 \cup F_2$ then $X_j = 10 w_i(i_j)$. Otherwise, $X_j = 0$. Let $W = \sum_{j=1}^{t} w_i(i_j)$. By the definition of t, $0.5 \leq W \leq 1$. Also, for every j, $w_i(i_j) < 0.1$ (otherwise, $i \in A_1$). We have $\mu = E\left[\sum_{j=1}^{t} X_j\right] \leq 10aW$. The event $i \in F_3$ happens only if $\sum_{i' \in (F_1 \cup F_2) \cap \{i_1, \ldots, i_t\}} w_i(i') < 0.1$, i.e. $\sum_{j=1}^{t} X_j > 10W - 1$. We bound the variance:

$$\begin{aligned} Var\left[\sum_j X_j\right] &= E\left[\left(\sum_j X_j\right)^2\right] - \left(E\left[\sum_j X_j\right]\right)^2 \\ &= \sum_j (10 w_i(i_j))^2 p_{i_j}(1 - p_{i_j}) \\ &+ 2 \sum_{j,j'} 100 w_i(i_j) w_i(i_{j'}) (Pr[(i_j \notin F_1 \cup F_2) \wedge (i_{j'} \notin F_1 \cup F_2)] - p_{i_j} p_{i_{j'}}) \end{aligned}$$

As $B \setminus (F_1 \cup F_2)$ is a set of fixed size, $Pr[(i_j \notin F_1 \cup F_2) \wedge (i_{j'} \notin F_1 \cup F_2)] \leq p_{i_j} p_{i_{j'}}$. Recall also that for each j, $10 w_i(i_j) \leq 1$. Therefore,

$$Var\left[\sum_j X_j\right] \leq \sum_j (10 w_i(i_j))^2 p_{i_j}(1 - p_{i_j}) \leq \sum_j 10 p_{i_j} w_i(i_j) = \mu$$

Using Chebyshev's inequality,

$$\begin{aligned} Pr\left[\sum_j X_j \geq 10W - 1\right] &= Pr\left[\sum_j X_j - \mu \geq 10W - 1 - \mu\right] \\ &\leq \frac{Var\left[\sum_j X_j\right]}{(10W - 1 - \mu)^2} \\ &\leq \frac{\mu}{(10W - 1 - \mu)^2} \end{aligned}$$

Note that $\mu \leq 10Wa \leq W$ (since $a \leq 0.1$). Also, $2W \geq 1$. Therefore,

$$Pr\left[\sum_j X_j \geq 10W - 1\right] \leq \frac{10Wa}{49W^2} \leq \frac{20}{49} a \leq \frac{1}{2} a$$ □

After determining the set of open facilities F', an optimal assignment of clients to facilities can be computed in polynomial time. However, it is easier to analyze a suboptimal assignment that is derived from the solutions S_1 and S_2. Consider a client $j \in C$. We distinguish between three cases:

Case 1: If $\varphi_2(j) \in F$, then we connect j to $\varphi_2(j)$. The connection cost is $d(j, \varphi_2(j))$, and this happens with probability at least b.

Case 2: If $\varphi_2(j) \notin F$, but $\varphi_1(j) \in F$, we connect j to $\varphi_1(j)$. The connection cost is $d(j, \varphi_1(j))$.

Case 3: Both $\varphi_1(j) \notin F$ and $\varphi_2(j) \notin F$. Put $i = \varphi_1(j)$. Define $U' = \{j' \in U_1(i) \mid \varphi_2(j') \in B'(i)\}$. Let $U'' = U_1(i) \setminus U'$. As $i \notin F_3$, $|U'| \geq 0.1|U_1(i)|$. Therefore, there is a function $g : U'' \to U'$, where at most 9 clients from U'' are mapped to each client in U'. We connect each client $j \in U''$ to $\varphi_2(g(j))$. The connection cost of j is bounded by $d(i) + d(i, j)$.

Lemma 10. If $a \leq 0.1$ then the above procedure generates a solution whose cost is at most $12z^*_{km}$.

Proof: The expected cost is bounded by:

$$\sum_{j \in C} (d(j, \varphi_2(j)) + a \cdot d(j, \varphi_1(j)) + a \cdot d(\varphi_1(j))) = \Phi(S_2) + a \cdot \Phi(S_1) + a \cdot \sum_{i \in A \setminus A_1} |U_1(i)| d(i)$$

We now bound the last term. Notice that by the definition of $d(i)$, for at least half the clients $j \in U_1(i)$, $d(i, \varphi_2(j)) \geq d(i)$. Therefore,

$$\sum_{i \in A \setminus A_1} |U_1(i)| d(i) \leq 2 \sum_{j \in C} d(\varphi_1(j), \varphi_2(j)) \leq 2 \sum_{j \in C} (d(j, \varphi_1(j)) + d(j, \varphi_2(j))) \leq 2\Phi(S_1) + 2\Phi(S_2).$$

As $a \leq 0.1$, $b \geq 0.9$. Thus, $2a \leq \frac{b}{2}$ and $1.5b > 1$. Therefore, the expected cost of the solution is at most:

$$\Phi(S_2) + a \cdot \Phi(S_1) + 2a \cdot \Phi(S_1) + 2a \cdot \Phi(S_2) \leq 3a \cdot \Phi(S_1) + 3b \cdot \Phi(S_2) \leq 12z^*_{km}. \qquad \square$$

Claim 11. The above procedure generates a solution where the capacities are exceeded by a factor of at most 50.

Proof: In the fractional solution S the capacities are violated by the factor of at most 5. For clients connected by case 1 or case 2, these are exactly their connections in solutions S_1 and S_2. So for each facility $i \in F$, there are at most $5u(i)$ such clients connected to it. For each such client, at most 9 additional clients are mapped to the same facility by the function g. Therefore, in total, for each facility i, at most $50u(i)$ clients connect to i. $\square$

The above discussion proves the main result of this paper:

Theorem 12. The algorithm presented in this section is a $(40, 50)$-approximation to the capacitated k-median problem.

References

[1] V. Arya, N. Garg, R. Khandekar, M. Meyerson, K. Mungala and V. Pandit. Local search heuristics for k-median and facility location problems. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 21-29, 2001.

[2] Y. Bartal, M. Charikar and D. Raz. Approximating min-sum k-clustering in metric spaces. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 11-20, 2001.

[3] M. Charikar and S. Guha. Improved combinatorial algorithms for facility location and k-median problems. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science,* pages 378-388, 1999.

[4] M. Charikar, S. Guha, É. Tardos, D.B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the* 31*st Annual ACM Symposium on Theory of Computing*, pages 1-10, 1999.

[5] F.A. Chudak. Improved approximation algorithms for uncapacitated facility location. In R.E. Bixby, E.A. Boyd, and R.Z. Rios-Mercado, editors, *Integer Programming and Combinatorial Optimization*, volume 1412 of *Lecture Notes in Computer Science*, pages 180-194. Springer, Berlin, 1998.

[6] F.A. Chudak and D.B. Shmoys. Improved approximation algorithms for the capacitated facility location problem. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 875-876, 1999.

[7] G. Cornuejols, G.L. Nemhauser and L.A. Wolsey. The uncapacitated faciilty location problem. In P. Mirchandani and R. Francis, editors, *Discrete Location Theory*, pages 119-171. John Wiley and Sonce Inc., 1990.

[8] S. Guha. Approximation algorithms for facility location problems. Ph.D. thesis, Stanford University, 2000.

[9] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms*, 31, pages 228-348, 1999.

[10] K. Jain, M. Mahdian, E. Markakis and A. Saberi. Greedy Facility Location Algorithms Analyzed using Dual Fitting. Submitted to *Journal of ACM*, 2002.

[11] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems, In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 731 - 740, 2002.

[12] K. Jain and V.V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *Journal of the ACM*, 48, pages 274-296, 2001.

[13] M.R. Korupolu, C.G. Plaxton and R. Rajaraman. Analysis of a local search heuristic for facility location problems. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1-10, 1998.

[14] J.-H. Lin and J.S. Vitter. ϵ-approximations with minimum packing constraint violation. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 771-782, 1992.

[15] M. Mahdian, E. Markakis, A. Saberi and V.V. Vazirani. A greedy facility location algorithm analyzed using dual fitting. In *Proceedings of 5th International Workshop on Randomization and Approximation Techniques in Computer Science,* volume 2129 of *Lecture Notes in Computer Science*, pages 127-137. Springer-Verlag, 2001.

[16] M. Mahdian, Y. Ye and J. Zhang. Improved approximation algorithms for metric facility location problems. In K. Jansen, S. Leonardi, V.V. Vazirani, editors, *Approximation Algorithms for Combinatorial Optimization*, volume 2462 of *Lecture Notes in Computer Science*, pages 229-242. Springer, 2002.

[17] A. Moss, Y. Rabani. Approximation algorithms for constrained node weighted steiner tree problems. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 373-382, 2001.

[18] M. Pál, É. Tardos, T. Wexler. Facility Location with Hard Capacities. In *Proceedings of the 42nd Annual IEEE Symposium on the Foundations of Computer Science*, pages 329-338, 2001.

[19] D.B. Shmoys. Approximation algorithms for facility location problems. In K. Jansen and S. Khuller, editors, *Approximation Algorithms for Combinatorial Optimization*, volume 1913 of *Lecture Notes in Computer Science*, pages 27-33. Springer, Berlin, 2000.

[20] D. B. Shmoys, É. Tardos and K.I. Aardal. Approximation algorithms for facility location problems. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, pages 265-274, 1997.

[21] M. Sviridenko. An improved approximation algorithm for the metric uncapacitated facility location problem. In W.J. Cook and A.S. Schulz, editors, *Integer Programming and Combinatorial Optimization*, volume 2337 of *Lecture Notes in Computer Science*, pages 240-257. Springer, Berlin, 2002.